Social network aggregation using face-recognition

Minder, Patrick ; Bernstein, Abraham

Abstract: With the rapid growth of the social web an increasing number of people started to replicate their off-line preferences and lives in an on-line environment. Consequently, the social web provides an enormous source for social network data, which can be used in both commercial and research applications. However, people often take part in multiple social network sites and, generally, they share only a selected amount of data to the audience of a specific platform. Consequently, the interlinkage of social graphs from different sources getting increasingly important for applications such as social network analysis, personalization, or recommender systems. This paper proposes a novel method to enhance available user re-identification systems for social network data aggregation based on face-recognition algorithms. Furthermore, the method is combined with traditional text-based approaches in order to attempt a counter-balancing of the weaknesses of both methods. Using two samples of real-world social networks (with 1610 and 1690 identities each) we show that even though a pure face-recognition based method gets outperformed by the traditional text-based method (area under the ROC curve 0.986 vs. 0.938) the combined method significantly outperforms both of these (0.998, p = 0.0001) suggesting that the face-based method indeed carries complimentary information to raw text attributes.

Originally published at:
Minder, Patrick; Bernstein, Abraham (2011). Social network aggregation using face-recognition. In: ISWC 2011 Workshop: Social Data on the Web, Bonn, Germany, 23 October 2011, RWTH Aachen.

Social Network Aggregation Using Face-Recognition

Patrick Minder and Abraham Bernstein

University of Zurich, Dynamic and Distributed Informations Systems Group
{minder,bernstein}@ifi.uzh.ch

Abstract. With the rapid growth of the social web an increasing number of people started to replicate their off-line preferences and lives in an on-line environment. Consequently, the social web provides an enormous source for social network data, which can be used in both commercial and research applications. However, people often take part in multiple social network sites and, generally, they share only a selected amount of data to the audience of a specific platform. Consequently, the interlinkage of social graphs from different sources getting increasingly important for applications such as social network analysis, personalization, or recommender systems. This paper proposes a novel method to enhance available user re-identification systems for social network data aggregation based on face-recognition algorithms. Furthermore, the method is combined with traditional text-based approaches in order to attempt a counter-balancing of the weaknesses of both methods. Using two samples of real-world social networks (with 1610 and 1690 identities each) we show that even though a pure face-recognition based method gets outperformed by the traditional text-based method (area under the ROC curve 0.986 vs. 0.938) the combined method significantly outperforms both of these (0.998, $p = 0.0001$) suggesting that the face-based method indeed carries complimentary information to raw text attributes.

1 Introduction

With the rapid growth of the social web an increasing number of people started to replicate their off-line preferences and lives in an on-line environment. Indeed, the usage of social network sites (SNS) such as Facebook, Google+ , or LinkedIN the use of messaging services (e.g., Twitter), tagging systems (e.g., del.ico.us), sharing and recommendation services (e.g., Last.fm) has not only increased immensely, but the activities on these site become an integral element in the daily lives of millions of people. Hence, the social web provides an enormous source for social network data collection.

Often people take part in multiple of these SNSs. In some cases this multi-participation arises from necessity, as some features may only be provided by some sites and not by others. However, in most cases, it is also the result of free choice. The many services allow people to "partition" their lives (e.g, they may

use facebook for the private- and LinkedIN for the professional network). In fact, the construction of site-specific identities enables the possibility to gain multiple personalities as identifying features, such as the email address can be changed easily—an effect that has been called "multiplicity" by Internet researchers [21]. Hence, users will continue to maintain multiple identities even if one service will cater to all their needs.

At the same time, the identification of users for interlinking data from different and distributed systems is getting increasingly important for different kind of applications. In personalization, the use of cross-site profiles is essential as the incorporation of multi-source user profile data significantly increases the quality of preference recommendations [4]; In social network analysis, the merging of multiple networks provides a more complete picture of the overall social graph and helps to minimize the data selection bias on which most single-site studies suffer [1]; and trust networks can be created by aggregating relationships among network participants [17]. Even if the semantic web were to become immensely popular the increased usage of a global identifier may not simplify universal identification of a person, as some sites may not use the same identifiers or even totally ignore the identification scheme and the users may choose—to ensure their multiplicity—to maintain multiple identifiers. In fact, Mika et al. [16] argue that the key problem in the area of extraction of social network data—the disambiguation of identities and relationships—still remains, as different social web applications refer to relationship types, attributes, or tastes in profiles in different ways and do not share any common key for the identification of users. As a consequence, both researchers and practitioners (such as marketers) are placed in front of a complicated research question: *how can we combine the multitude of information available about a person in the multiple SNSs to develop a holistic, combined (and as complete as possible) user model when the identity of the user in different sites is difficult to combine?*

Current proposals for interlinking social network profiles based on comparing text-based attributes of user profiles [4] or using the network structure [13] have the drawback that these methods scale poorly or they need to contain some overlap in the relationship structure and result in a large computational expenditure respectively. In this paper we propose to enhance current text-based methods—in absence of semantic metadata — by combining it with face recognition algorithms. Specifically, we propose to use face-recognition software to compare the images uploaded by users on different SNSs as an additional feature for identity merging. As we show, this statistical entity resolution procedure significantly enhances the merging precision of two SNSs. Consequently, *the contribution of this paper are: (1)The presentation of an enhanced identity merging framework to incorporate images; (2) The presentation of an algorithm that merges identities based on face recognition software. (3) The combination of traditional text-based and the introduced image-based merge-approach to counter-balance the respective weaknesses of each of the approaches.*

To this end, we first ground our idea by giving an overview of related work and introducing the fundamental concepts of entity resolution (i.e. re-identification)

and face-recognition. Then we present our novel re-identification technique and discuss our prototype. Finally, we evaluate our procedure empirically on three real-world datasets and close with a discussion of the limitations, future work and some general conclusions.

2 Related Work

Winkler [26], showed that with a minimal set of attributes a large portion of the US population can be re-identified based on US Census data. Furthermore, Gross et al. [10] showed that about 80% of social network sites user provide enough public data for a direct re-identification and that at least 61% of the published profile images on Facebook.com allow a direct identification by a human.

Carmagnola et al.[4] and Bekkermann et al. [2] provide a cross-system identity discovery system, which is based on text-based identification probability calculations, whereby public available textual attributes of social network sites are analyzed by their positive, respectively negative, influence on identification. Further, [3] suggest the use of key phrase extraction for the name disambiguation process, which is also used in POLYPHONET [14] for interlinking web pages

[13] and [22] provide re-identification algorithms based on network similarity. These system provide high accuracy, but lack on computational complexity and time expenditure.

A lot of research concerns shared approaches [12]: Especially, the application of common semantic languages , such as the FOAF ontology[1], the SIOC (Semantically-Interlinked Online Communities) ontology[2] for online communities or the SCOT (Social Semantic Cloud Of Tags) ontology[3] for tagging systems. Such systems are desirable, but not widely spread in reality. The most well-known system based on such data is FLINK [15].

3 Theoretical Foundations

In this section, we present the theoretical foundations for our approach. First, we present a formal model for entity resolution and then succinctly explain the basics of face-recognition. Both foundations are used in our framework.

3.1 Entity Resolution and the Fellegi-Sunter Model

Entity resolution can be defined as *the methodology of merging corresponding records from two or more sources* [26]. Consider for example a profile about "Peter J. Miller" and another one about "Peter Jonathan Miller" on two different SNS. Entity Resolution tries to decide if these two profiles belong to the same

[1] http://www.foaf-project.org / http://xmlns.com/foaf/spec/20100101.html
[2] http://sioc-project.org/
[3] http://scot-project.org/

person or not. Therefore, entity resolution assumes that an individual shares similar features in different environments which can be used to identify an entity, even though no common key is defined. Generally, to complicate the resolution process, there are different entities that share similar attribute values.

Most current re-identification approaches are variants of the Fellegi-Sunter model—a distance- and rule-based technique. The Fellegi-Sunter Model determines a match between two entities by computing the similarity of their attribute (or feature) vectors [9]. Specifically, given entities $a \in \mathbb{A}$ and $b \in \mathbb{B}$, where both $\mathbb{A}$ and $\mathbb{B}$ are the set of entities in SNS A and B, it tries to assign each pair (a, b) of the space $\mathbb{A} \times \mathbb{B}$ to a set $\mathbb{M}$ or $\mathbb{U}$ whereby:

$$\mathbb{M} := \text{is the set of true matches} = \{(a, b); a \in \mathbb{A} \wedge b \in \mathbb{B} \wedge a = b\}$$
$$\mathbb{U} := \text{is the set of non-matches} = \{(a, b); a \in \mathbb{A} \wedge b \in \mathbb{B} \wedge a \neq b\}$$

It does so using a comparison function γ that computes the similarity measures for each of the n comparable attributes of the entities and arranges these in a vector:

$$\gamma(a, b) = \{\gamma^1(a, b), ..., \gamma^n(a, b)\}$$

Based on the comparison vector $\gamma(a, b)$ a decision rule L now assigns each pair (a, b) to either to the set M or U as follows:

$$(a, b) \in \begin{cases} M & \text{if } p(M|\gamma) \geq p(U|\gamma) \\ U & \text{otherwise} \end{cases}$$

whereby $p(M|\gamma)$ is the probability that the comparison vector γ belongs to the match class and $p(U|\gamma)$ that γ belongs to U. In other words, the Fellegi-Sunter model treats all pairs of possible matches as independent. Recently several authors argued that this independence offers the opportunity for enhancements. Singla et al [18], e.g., proposes such an enhancement based on Markov logic.

3.2 Face-Recognition and the Eigenface Algorithm

The face provides an enormous set of characteristics that the human perception system uses to identify other individuals. The problem of face-recognition can be formulated as follows *"Given still or video images of a scene, identify or verify one or more person in the scene using a stored database of faces. Available collateral information [...] may be used in narrowing the search (enhancing recognition)"* [25, p. 4]. Accordingly, face-recognition includes [25]: *(1)* The detection and location of an unknown number of faces in an image [11]; *(2)* The extraction of key facial-features; and *(3)* The identification [25, p. 12] which includes a comparison and matching of invariant biometric face signatures [25, p. 14 - 16]. The identification can either be done by using *holistic matching*, *feature-based matching*, or *hybrid matching methods* which concern the whole face, local features— e.g. the location or geometry of the nose —or both as an input vector for classification respectively [25, p. 14].

Our re-identification framework uses the holistic face-recognition algorithm *Eigenface* [20] based on Principal Component Analysis (PCA) and covering all relevant local and global features [20]. The Eigenface approach tries to code all the relevant extracted information of a face image, such that the encoding can be done efficiently, allowing for a comparison of the information to a database of encoded models [25, p. 67].The Eigenface algorithm can be split up into two parts:

(1) Representation of the Image Database in Principal Component Vectors Based on PCA, the principal components of a face-image are extracted, by (1) acquiring an initial set of face images; (2) Defining the face space by calculating the eigenvectors (Eigenfaces) from the set and eliminating all but k best eigenvectors with the highest eigenvalues, by using PCA; and (3) Presenting each known individual by projecting their face image onto the face space.

Therefore, an image $I(x, y)$ can be interpreted as a vector in a N-dimensional space, where $N = rc$ and r are the rows and c columns of the image [20] . Every coordinate in the N-dimensional vector $I(x, y)$— the *image space* —corresponds to a pixel of the image. This representation of an image obfuscates any relationship between neighboured pixels as long as all images are rearranged in the same manner. Thus the average face of the initially acquired training set $\Gamma := \{\gamma_1, \gamma_2, ..., \gamma_m\}$ can be calculated by

$$\bar{\gamma} = \frac{1}{m} \sum_{n=1}^{m} \gamma_n.$$

and the distance between an image and the average image is measured by $\varphi_i = \gamma_i - \bar{\gamma}$. Whereby, the orthonormal vectors define an Eigenface with the eigenvectors:

$$u_l = \sum_{k=1}^{M} e_{lk} \varphi_k \forall i \in [1, M]$$

whereby the eigenvectors e_l are calculated from the covariance matrix $L = AA^\top$, where $L_{mn} = \varphi_m{}^\top \varphi_n$ and $A = [\varphi_1, \varphi_2, ..., \varphi_M]$. The derivation of the best eigenvectors out of the covariance matrix is presented in [19]. The k significant eigenvectors of L span an k-dimensional face space—a subspace of the $N \times N$ dimensional image space—where every face is represented as a linear combination of the Eigenfaces [20] [25, p. 67 - 72].

(2) The Identification Process The identification respectively verification of an image is processed by: (1)Subtracting the mean image from the new face images and projecting the result onto each of the eigenvectors (Eigenfaces); (2) Determining if the image is a face by calculating the distance to the face space and comparing it to a defined threshold; and (3)If it is a face, classifying the weight pattern as a known or unknown individual by using a distance metric, such as the Euclidian distance.

Thus, a new face image $I(x, y)$ will be projected into the face space by $\omega_k = u_k^\top(\gamma - \bar{\gamma})$ for $\forall k = [1, ..., M']$. The weight matrix $^\top = [\omega_1, ..., \omega_{M'}]$ represents the influence of each eigenvector on the input image. Hence, given a threshold θ_ε, if the face class k, which minimizes the Euclidian distance is

$$\varepsilon_k = \| (\quad - \quad_k) \| \text{ and } \theta_\varepsilon > \varepsilon_k \tag{1}$$

then the image will belong to the same individual. Else the face is classified as unknown. Furthermore, the distance between an image and the face space can be characterised by the squared distance between the mean-adjusted input image:

$$\varepsilon^2 = \| (\varphi - \varphi_f) \| \text{ , where } \varphi = \gamma_k - \bar{\gamma} \text{ and } \varphi_f = \sum_{i=1}^{M'} \omega_i u_i \tag{2}$$

Therefore, a new face image $I(x, y)$ will be calculated as a non-face image if $\varepsilon > \theta_\varepsilon$, as known face image if $\varepsilon < \theta_\varepsilon \wedge \varepsilon_k < \theta_\varepsilon$ and as an unknown face image if $\varepsilon < \theta_\varepsilon \wedge \varepsilon_k > \theta_\varepsilon$.

4 Re-Identification Framework

Our theoretical re-identification framework for user disambiguation in a social network aggregation and cross-system personalization process. is based on the Fellegi-Sunter-Model. The presented algorithms calculate the probability that two user profiles belong to the same entity, and incorporates the ability to incorporate images as an additional feature based on the Eigenface method. Therefore, the framework provides three kind of methods: a pure face-recognition based, a text-attribute based, and joined re-identification method.

The methods follow a simple re-identification algorithm. Assume, two sets $\mathbb{A} = \{a_1, a_2, ..., a_m\}$ and $\mathbb{B} = \{b_1, b_2, ..., b_n\}$ of user profiles from two different SNSs. Each profile is characterized by a set of text attributes and a single profile image. We can now define $\mathbb{E} = \{e_1, e_2, ..., e_z\}$ as the set of different individuals, who have a profile in one or both social networks. Consequently, the re-identification algorithm is based on the following three subtasks:

1. *Attribute Comparison:* The attributes of two social network profiles are compared pairwise. The result is a comparison vector $\gamma(a_i, b_j) = \{d_1, d_2, ..., d_n\}$, where n is the number of compared attributes and $d_k \in [0, 1]$ indicates the distance between the values of the k^{th}-attribute of the profiles a_i and b_j. Therefore, a distance d_k of 0 indicates, that the two attribute instances are completely equal, and a value of 1 indicates the opposite.
2. *Matching Probability Calculation:* Then, based on the comparison vector $\gamma(a_i, b_j)$, the probability $\rho(a_i, b_j)$, that a pair (a_i, b_j) belongs to the same entity, is calculated.
3. *Merging Task:* Finally, if probability $\rho(a_i, b_j)$ is greater or equal to a threshold value $\theta \in [0, 1]$ (*i.e.*, $\theta \geq \rho(a_i, b_j)$) then the profiles a_i and b_j are assumed to belong to the same person.

4.1 Attribute Comparison and Matching Probability Calculation

The following three generic methods allow the comparison of n different attributes and the calculation of a matching probability. The methods cover the

first two subtasks of the above introduced re-identification algorithm.

(1) Pure Face-Recognition Based Method The method re-identifies user profiles only by the application of the face-recognition algorithm $Eigenface$ on profile images. Hence, $\forall a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$, the probability $\rho(a_i, b_j)$, that two profiles a_i and b_j belong to the same entity $e_l \in E$, is defined as:

$$\rho(a_i, b_j) = \varepsilon_{ij}(a_i, b_j) = \| (\quad_{a_i} - \quad_{b_j}) \| \in [0, 1]$$

Whereas, it is assumed that the profile images are projected into the face space by $\omega_{a_i} = u_k^\top(a_i - \bar{\gamma})$ and $\omega_{b_j} = u_k^\top(b_j - \bar{\gamma})$. Additionally, the set $\mathbb{B}$ is used as training set for the initialization task, thus $\Gamma = \mathbb{B}$.

(2) Text-Attribute Based Method The algorithm re-identifies user profiles by the application of text-attribute comparison. The attributes are compared with the token-based $QGRAM$ algorithm [7]. Note that spelling errors minimally affects the similarity when using $QGRAM$, as it uses q-grams instead of words are used as tokens. For the k^{th}-attribute the algorithm computes a normalized distance $d(a_{ik}, b_{jk}) \in [0, 1]$, where the distance is zero, if the value of the k^{th}-attribute of a_i and b_j are syntactically equivalent. As we discuss in Section 6, we considered $name$, $email\ address$, $birthday$, $city$ as a minimal set of text attributes in the experiments as they where shown to be strong indicators for identification [5] [26] [10] and other attributes such as address or phone number are often not accessible. As a result, the matching probability is calculated by a logistic function [8]:

$$\rho(a_i, b_j) = \frac{\exp(Y_T(a_i, b_j))}{1 + \exp(Y_T(a_i, b_j))} \in [0, 1]$$

where

$$Y_T(a_i, b_j) = \alpha_0 + \sum_{k=1}^{n} \alpha_k d(a_{ik}, b_{jk})$$

The intercept α_0 and regression coefficients $\{\alpha_1, ..., \alpha_n\}$ for the linear regression model $Y_T(a_i, b_j)$ are learned by a logistic regression on a specific training set.

(3) Joined Method Finally, the two previously described methods are joined to a method that uses both face-image-based and text-attribute-based identification. Thus, for all pairs of profiles $a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$, it is assumed that the matching probability is equal to:

$$\rho(a_i, b_j) = \frac{\exp(Y_J(a_i, b_j))}{1 + \exp(Y_J(a_i, b_j))} \in [0, 1]$$

where

$$Y_J(a_i, b_j) = \alpha_0 + \sum_{k=1}^{n} \alpha_k d(a_{ik}, b_{jk}) + \beta \varepsilon_{ij}(a_i, b_j)$$

Again, the intercept α_0 and regression coefficients $\{\alpha_1, ..., \alpha_n, \beta\}$ for the linear regression model $Y_J(a_i, b_j)$ are learned by a logistic regression on a specific training set.

4.2 Merging Task

Finally, based on one of the above introduced matching probabilities, a pair (a_i, b_j) is called to belong to the same entity (*i.e.*, $(a_i, b_j) \in \mathbb{M}$), if:

$$\forall a_i \in A \wedge b_j \in B : \theta \geq \rho(a_i, b_j) \longrightarrow \mathbb{M} \tag{3}$$

5 Prototype

Our re-identification framework consists of four major components. Currently, the *Data Gathering and Acquisition* module enables the acquisition of network data from the social network sites Facebook, LinkedIn, Twitter and Flickr, whereby only concerns public available data. The *Data Preprocessing* module preprocesses the crawled data by transforming profile attributes into an internal schema and establish connections between profiles for each relationship in the source network. The implementation provides functionality for both the integration of text attributes and profile images. For the integration of profile images, we use an implementation of the face detection algorithm *OpenCV*[4] *HaarClasifier*[23] provided by the Faint[5] open source project. The algorithm returns the coordinates of every face region on an input image, whereby one region of the n returned regions is randomly selected and resized to a 50×50-pixel image. The *Matching* module performs a pairwise comparison of all possible profiles pairs (a_i, b_j), where $a_i \in \mathbb{A} \wedge b_j \in \mathbb{B}$. The goal of the matching task is to calculate the comparison vector $\gamma(a_i, b_j)$ and matching probability $\rho(a_i, b_j)$ for each of the methods introduced in Section 4.1. The module uses text-based algorithm QGRAM provided by the open-source project SimMetrics[6], and our own implementation of the Eigenface algorithm. Finally, The *Merging* module merges the data sources to an aggregated social graph based on rule introduced in Section 4.2.

6 Experiments

We evaluated the accuracy of the framework based on two experiments. In the first experiment we determined various input parameters, the intercept and the coefficients for the two regression models. The second experiment benchmarked the suitability of profile images for user disambiguation in the pure face-recognition and joined method against the text-based matching algorithm.

6.1 Experiment 1: Determining the Parameters

In the first experiment two social networks with a size of 47 and 45 were generated from data crawled on Facebook. 36 of these users had a profile in both

[4] http://sourceforge.net/projects/opencvlibrary/
[5] http://faint.sourceforge.net/
[6] http://www.sourceforge.net/projects/simmetrics/

networks. The profile image was randomly selected from all public available published images in the specific Facebook profile. We performed a pairwise comparison of the two sets, whereas for each pair the attribute similarities were stored as a quintuple [name, emailaddress, birthday, city, $\text{image}_{similarity}$] whilst varying the number of Eigenfaces in the $\text{image}_{similarity}$ computation. Finally, the optimal number of Eigenfaces and parameters for the two linear models were calculated using a logistic regression model in SPSS[7].

Performance metric The profile image similarity measurements based on Eigenfaces were compared using Receiver Operating Characteristics (ROC) curves. The ROC-curve graphs the true positive rate (y-axis) respectively sensitivity against the false positive rate (x-axis) respectively 1 - Specificity, where an ideal curve would go from the origin (0,0) to the top left (0,1) corner, before proceeding to the top right (1,1) one [24, p. 244 - 225]. The area under the ROC-curve (AUC, also called c-statistic in medicine) can be used as a single number performance metric for the merge accuracy. In contrast to the traditional precision, recall, or f-measure it has the advantage that both the ROC-curve and the AUC are independent of the prior data-distribution and, hence, serve as a more robust metric to compare the performance of two approaches.

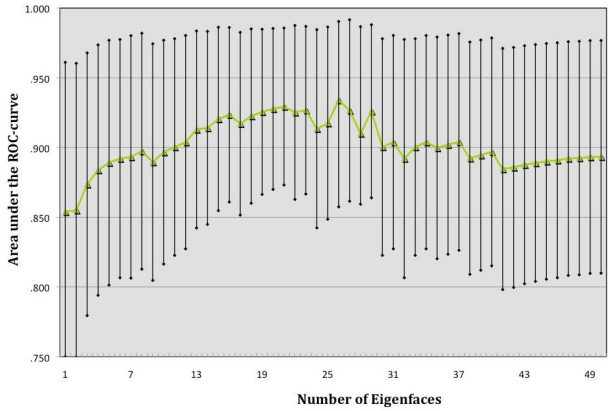

Fig. 1: Showing the influence of the number of Eigenfaces on the area under the ROC-Curve based on data of the first experiment and a confidence interval of 95%

Results As illustrated in Figure 1, the number of Eigenfaces influences on the accuracy of match. The accuracy of the algorithm increases when increasing the number of Eigenfaces until a specific barrier, where any increase in its numbers is not beneficial or even detrimental to the overall performance. Thus, the Eigenface algorithm should use between 50 to 60% of the top-most Eigenfaces—a result similar to [24]. The resulting input parameters for the linear models are shown in Table 1.

Computational Costs The computational costs for the face-image comparison is higher than for single text-based comparison. On our test-machine (an

[7] http://www.spss.com/

Apple iMac computer with a 3.06 GHz Intel Core 2 Duo processor and 4 GB of RAM) the comparison of the four concerned text-attributes takes between 10 to 20ms per pair without data preprocessing; the image-based comparison alone takes 25 to 35ms/pair. Additionally, once per image, the face preprocessing, including face-detection and image resizing, takes between five and six seconds.

Attribute	α_0	α_{Name}	α_{Email}	$\alpha_{Birthday}$	α_{City}	β
Text-Based Method Y_T	-0.319	25.655	-1.763	9.750	25.334	-
Joined Method Y_J	-6.659	26.656	0.234	11.536	18.272	8.788

Table 1: Input parameter for the regression based text-based and joined method models learned on the dataset of the first experiment and used in the second experiment as input.

6.2 Experiment 2

For the second experiment we collected a subgraph of both Facebook and LinkedIn. Departing from the first author's profile we collected 1610 (Facebook) respectively 1690 (LinkedIn) profiles and manually determined that 166 users where present in both samples. We compared all these profiles with the three approaches using the input parameters determined in Experiment 1. **Results** Figure 2 graphs the ROC curves for the three methods. Note that whilst the text method (AUC= 0.986) outperforms the pure image-based method (AUC= 0.938), the combined method (AUC= 0.998) significantly outperforms either methods ($p = 0.001$, $p = 0.0001$ compared with a non-parametric method described by DeLong [6]).

6.3 Discussion, Limitations and Future Work

As the above results show the combined method clearly outperforms each of others. It is interesting to observe that the ROC-Curve of both text-based and the image-based method both shoot almost straight up until about (0,0.9). Then the text-based method flattens out whilst the combined one continues to rise. This suggests that the element of the method's accuracy is contributed mostly by the image-based method. Only then does the image-based method contribute additional predictive power. When looking at the regression parameters this suggestion receives some additional support as the parameters for the Email and City lose in their contribution whilst the algorithm relies more on the Name, Image, and interestingly the Birthday.

Obviously, all these results are limited by the usage of only one, albeit real-world, dataset and will have to be validated with a others. Also, our experiment assumed that we knew the semantic alignment of the text-attributes. When merging only two SNS this assumption seems reasonable, when more are involved the this alignment may introduce additional error. Consequently, we probably overestimated the accuracy of the textual method.

Last but not least, a real-world system would probably not perform a full pairwise comparison to limit the computational expenditure but use some optimization approach.

We intend to investigate all these limitations in our future work.

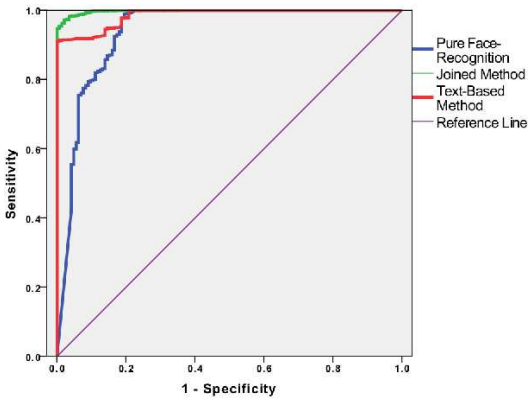

Fig. 2: Results of the second experiment merging two subnetworks of Facebook and LinkedIn

7 Discussion and Conclusion

In this paper we proposed an extension of the traditional text-attribute-based method for re-identification in social networks using the images of profiles. The experimental results show that the pure face-recognition based re-identification method does not compete the traditional text-based methods in accuracy and computational performance. A combined method, however, significantly outperforms the pure text-based method in accuracy suggesting that it contains complementary information. As we showed this combined method significantly improves the accuracy of a social network system merge. Consequently, we believe that it provides a more solid basis for both researchers and practitioners interested in investigating multiple SNSs and facing the problems of multiplicity.

References

1. Bachmann, A., Bird, C., Rahman, F., Devanbu, P., Bernstein, A.: The missing links: Bugs and bug-fix commits. In: ACM SIGSOFT / FSE '10: Proceedings of the 18th International Symposium on the Foundations of Software Engineering (2010)
2. Bekkerman, R., McCallum, A.: Disambiguating web appearnaces of people in a social network. In: Procceding of the WWW 2005 (2005)
3. Bollegara, D., Matsuo, Y., Ishizuka, M.: Extracting key phrases to disambiguate personal names on the web. In: Proceeding to CICling 2006 (2006)
4. Carmagnola, F., Cena, F.: User identification for cross-system personalisation. Information Sciences: an International Journal 1-2(179), 16–32 (2009)
5. Carmagnola, F., Osborne, F., Torre, I.: User data distributed on the social web: how to identify users on different social systems and collecting data about them. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (2010)
6. DeLong, E., DeLong, D., Clarke-Pearson, D.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44(44), 837 – 845 (1988)

7. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1) (2007)
8. Fahrmeir, L., Pigeot, I., Tutz, G.: In: Statistik - Der Weg zur Datenanalyse. Springer-Verlag Berlin Heidelberg New York (2003)
9. Fellegi, I., Sunter, A.: A theory for record linkage. Journal American Statistic Association (64), 1183 – 1210 (1969)
10. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Workshop On Privacy In The Electronic Society, Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society. pp. 71 – 80 (2005)
11. H Demirel, TJ Clarke, P.C.: Adaptive automatic facial feature segmentation. Proc. of 2nd International Conference on Automatic Face and Gesture Recognition pp. 277 – 282 (196)
12. Leonard, E., Houben, G.J., van der Sluijs, K., Hidders, J., Herder, E., Abel, F., Krause, D., Heckmann, D.: User profile elicitation and conversion in a mashup environment. In: Int. Workshop on Lightweight Integration on the Web, in conjunction with ICWE 2009 (2009)
13. Malin, B.: Unsupervised Name Disambiguation via Social Network Similarity (2006)
14. Matsuo, Y., Mori, J., Hamasaki, M., Ishizuka, M.: Polyphonet: An advanced social network extraction system. In: Procceding of 15th International World Wide Web Conference (2006)
15. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics 3(211 - 223) (Jan 2005)
16. Mika, P., Gangemi, A.: Descriptions of social relations. In: Proceedings of the First Workshop on Friend of a Friend, Social Network and the Semantic Web (2004)
17. Rowe, M.: Interlinking distributed social graphs. In: Proc. Linked Data on the Web Workshop, 18th Int. World Wide Web Conference (2009)
18. Singla, P., Domingos, P.: Entity resolution with markov logic. In: ICDM Sixth International Conference on Data Mining 2006 (2006)
19. Sirovich, L., Kirby, M.: Application of the karhunen-loève procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence . . . 12, 103 – 108 (1990)
20. Turk, M., Pentland, A.: Face recognition using eigenfaces. Conference on Computer Vision and Pattern Recognition (Jan 1991)
21. Turkle, S.: Cyberspace and identity. Contemporary Sociology 28(6), 643 – 648 (1999)
22. Veldman, I.: Matching Profiles from Social Network Sites. Master's thesis, University Twente (2009)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recogntiion (2001)
24. Wechsler, H.: Reliable Face Recognition Methods - System, Design, Implementation and Evaluation. Springer Media LLC (2"6)
25. Wenyi Zhao, R.C.: Face Processing - Advanced Modeling and Methods. Academic Press (2006)
26. Winkler, W.: The state of record linkage and current research problems. Tech. rep., Statistical Research Division, U.S. Census Bureau. (1999)